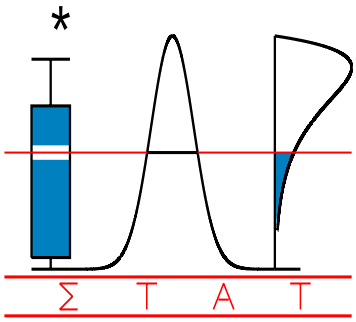
T E C H N I C A L

R E P O R T

0407

COMPARISON OF DIFFERENT ESTIMATION PROCEDURES FOR PROPORTIONAL HAZARDS MODEL WITH RANDOM EFFECTS

José CORTINAS ABRAHANTAS, Catherine LEGRAND, Tomasz BURZYKOWSKI, Paul JANSSEN, Vincent DUCROCQ and Luc DUCHATEAU



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Comparison of Different Estimation Procedures for Proportional Hazards Model with Random Effects

José Cortiñas Abrahantes [a,*] Catherine Legrand [b]
Tomasz Burzykowski [a] Paul Janssen [a] Vincent Ducrocq [c]
Luc Duchateau [d]

[a] *Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B3590 Diepenbeek, Belgium.*

[b] *European Organization for Research and Treatment of Cancer (EORTC), B1200 Brussels, Belgium.*

[c] *Station de Génétique Quantitative et Appliquée, Institut National de la Recherche Agronomique,78352 Jouy en Josas, France.*

[d] *Faculty of Veterinary Medicine, Department of Physiology, Biochemistry and Biometrics, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium.*

Abstract

Proportional hazards models with multivariate random effects (frailties) acting multiplicatively on the baseline hazard is a topic of intensive research. Several estimation procedures have been proposed to deal with these type of models. In this paper four estimation procedures used to fit these models (McGilchrist and Aisbett, 1991; Ducrocq and Casella, 1996; Ripatti and Palmgren, 2000; Cortiñas and Burzykowski , 2004) are compared in a simulation study. The performance of the four methods are compared based on their point estimates and the standard error associated to the estimates. From the simulation study can be concluded that McGilchrist and Aisbett approach face problems with the estimation of the standard error of the variance parameters, while the other three methods produce comparable results.

Key words: Frailty model; Residual maximum likelihood; Penalized partial likelihood; Laplace approximation; Multivariate failure-time data.

* Corresponding author. Tel.: +32-11-268215; fax: +32-11-268299
Email address: jose.cortinas@luc.ac.be (José Cortiñas Abrahantes).

1 Introduction

In applied sciences, one is often confronted with the collection of ***correlated data***. This generic term embraces a multitude of data structures, such as multivariate observations, clustered data, repeated measurements, longitudinal data, and spatially correlated data. Instances of this type of research can be encountered in virtually every empirical branch of science.

In this paper we will focus on clustered failure-time data. A way to model the data is to use a proportional hazards model conditional on random effects introduced to allow for correlation between the observations from the same cluster. First proposals for such a modelling strategy concentrated on the univariate mixed effects model (also called shared frailty model), which only include a univariate random effect in the model.

However, shared frailty models have some limitations. For instance, they force the unobserved factors (frailty) to be the same for all failure-times within the cluster (Xue and Brookmeyer, 1996). This may not always be desirable. Another drawback is that in most cases, a univariate frailty can only induce positive association within the cluster (Xue and Brookmeyer, 1996). Clearly, there are some situations in which the failure-times for subjects within the same cluster may be negatively associated.

To avoid the limitations, models with multivariate, correlated random effects have been proposed. The main problem with the use of such models is the estimation of their parameters. Several estimation approaches have been proposed. McGilchrist and Aisbett (1991), McGilchrist (1993) and McGilchrist (1994) used a penalized likelihood approach. Xue and Brookmeyer (1996) proposed the EM algorithm with numerical integration used at the E-step. Ducrocq and Casella (1996) developed a Bayesian approach. Vaida and Xu (2000) used the Monte Carlo EM (MCEM) algorithm, with Monte Carlo Markov Chain (MCMC) sampling used at the E-step. Ripatti, Larsen and Palmgren (2002), following the ideas of Vaida and Xu (2000), introduced an estimation procedure based on the MCEM algorithm, in which they used rejection sampling to draw from a posterior distribution of the random effects at the E-step. Cortiñas and Burzykowski (2004) proposed a modification of the EM algorithm in which the Laplace approximation is used at the E-step. Xue (1998) developed an alternative fitting method using estimating equations derived from a Poisson regression formulation, while Xue and Ding (1999) used a Gibbs sampling approach. Ripatti and Palmgren (2000) proposed estimation based on a penalized partial likelihood developed by applying the Laplace approximation to the marginal likelihood function.

The aim of this paper is to compare different estimation procedures used to

fit proportional hazards models with random effects. To this aim, a simulation study is conducted. We consider the methods developed by McGilchrist and Aisbett (1991), Ducrocq and Casella (1996), Ripatti and Palmgren (2000) and Cortiñas and Burzykowski (2004). The main reason for this choice was software availability and numerical complexity. The paper is organized as follows. Section 2 briefly recalls the proportional hazards model with random effects. In Section 3 the four estimation methods are reviewed. Section 4 describes the simulation study. The results are presented in Section 5. A short discussion of the results in Section 6 concludes the paper.

2 The Proportional Hazards Model with Random Effects

We consider clustered failure-time data with N clusters. The failure-time variable corresponding to subject j ($j = 1, \ldots, n_i$) from cluster i ($i = 1, \ldots, N$) will be denoted by Y_{ij}. It is assumed that observations of Y_{ij} can be right-censored. Thus, for subject j in cluster i we observe $T_{ij} = \min(C_{ij}, Y_{ij})$, where C_{ij} is a censoring time independent of Y_{ij}. Additionally, a censoring indicator δ_{ij} is observed, with δ_{ij} equal to 1 if $T_{ij} = Y_{ij}$, and 0 otherwise.

As we mentioned in the introduction the univariate shared frailty model was the first proposal to handle clustered failure-times data. It can be written as

$$\lambda(t_{ij}|\beta, \omega_i) = \lambda_0(t_{ij})\omega_i \exp(x_{ij}^T\beta), \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function, β is a vector of fixed-effects corresponding to a vector of covariates x_{ij}, and cluster-specific random effects ω_i are assumed to be independent, identically distributed random variables with a common density function $f(\omega_i; \theta)$, where θ is the parameter quantifying the variability of frailties. One of the most common distribution assumed for the frailties is the gamma distribution (Clayton, 1978; Vaupel, Manton and Stallard, 1979; Oakes, 1982; Hougaard, 2000). The main reason is that in this case it is easy to derive closed form expressions of marginal survival, density and the hazard. In the case of a parametric hazard, if the random effects are gamma distributed, analytic expression for the likelihood can be derived. On the other hand, if the hazard is unspecified, then EM algorithm with closed form expression for the conditional expectation of the frailties can be used. It is worth noting that model (1) can be rewritten in the following form:

$$\lambda(t_{ij}|\beta, b_i) = \lambda_0(t_{ij}) \exp(x_{ij}^T\beta + b_{i0}), \tag{2}$$

where $b_{i0} = \ln \omega_i$. In what follow, we will distinguish between "frailties" ω_i and "random effects" b_i.

In this paper we consider the following extension of model (2), which can be written as

$$\lambda(t_{ij}|\beta, b) = \lambda_0(t_{ij}) \exp(x_{ij}^T\beta + z_{ij}^T b_i), \tag{3}$$

where $\lambda_0(t)$ and β are the baseline hazard function and the vector of fixed effects, respectively, and b_i is a d-dimensional vector of random effects associated with a vector of covariates z_{ij}. We will assume that the random effects $b_i^T = (b_{i0}, b_{i1}, ..., b_{id})$ are normally distributed with mean 0 and variance-covariance matrix $D = D(\theta)$. To simplify formulas, we will also use the baseline cumulative hazard defined as

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du.$$

Model (3) can be seen as a linear mixed-effects model on the log-hazard scale. The estimation of the parameters β and θ from the observed data on T_{ij} is our main interest. Assuming the conditional independence of the observations within a cluster given b_i, one might write the (conditional) log-likelihood for the observed data as

$$l^C(\beta, \lambda_0, b) = \sum_{i=1}^{N} l_i^C(\beta, \lambda_0, b_i), \tag{4}$$

where

$$l_i^C(\beta, \lambda_0, b_i) = \sum_{j=1}^{n_i} [\delta_{ij}\{\ln \lambda_0(t_{ij}) + x_{ij}^T\beta + z_{ij}^T b_i\} - \Lambda_0(t_{ij}) \exp(x_{ij}^T\beta + z_{ij}^T b_i)] \tag{5}$$

is the (conditional) log-likelihood for the observed data in the ith cluster, and b denotes the vector resulting from "stacking" vectors b_i for all clusters. The (marginal) likelihood of the observed data for all clusters can then be expressed as

$$L^M(\beta, \theta, \lambda_0) = \prod_{i=1}^{N} \int L_i^A(\beta, \theta, \lambda_0, b_i)db_i, \tag{6}$$

where

$$L_i^A(\beta, \theta, \lambda_0, b_i) = f(b_i; \theta) \prod_{j=1}^{n_i} e^{l_i^C(\beta, \lambda_0, b_i)}. \tag{7}$$

and $f(b_i;\theta)$ is the density function of b_i. Note that (7) can be treated as the likelihood of the "augmented" data for cluster i, treating b_i as additional observations. Consequently,

$$L^A(\beta,\theta,\lambda_0,b)=\prod_{i=1}^{N}L_i^A(\beta,\theta,\lambda_0,b_i), \tag{8}$$

is the likelihood of the "augmented" data for all clusters.

One might consider using directly the likelihood function (6) in the inference on β and θ. There are, however, two major problems with using it for this purpose. First, it depends on the baseline hazard function λ_0. Unless a parametric form of the hazard can be assumed, the usefulness of (6) is limited. Second, the integral in (6) will usually be multi-dimensional, unless a very simple model is considered, and in general will not be available in a closed form.

Several estimation approaches have been proposed to circumvent these problems. In the following section some of the methods are reviewed.

3 Estimation Methods

In this section the approaches proposed by McGilchrist and Aisbett (1991), Ducrocq and Casella (1996), Ripatti and Palmgren (2000) and Cortiñas and Burzykowski (2004) are reviewed. In what follows, we will assume that b_i are normally distributed with mean 0 and variance-covariance matrix

$$D(\theta)=\begin{pmatrix}\theta_0 & 0 & \dots & 0\\ 0 & \theta_2 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \theta_d\end{pmatrix}.$$

3.1 REML Estimation Method

McGilchrist and Aisbett (1991) used the penalized likelihood approach to estimate the fixed effects and the residual maximum likelihood (REML) to estimate the variance components of the random effects. Their method consists of finding the best linear unbiased predictors (BLUP) of the fixed and random

components in first place, and then to use them to find REML estimates of the variance covariance parameters.

3.1.1 BLUP and REML Estimators

The estimation procedure is a generalization of the results developed by Schall (1991). In order to find the BLUP it is necessary to maximize the sum of two components. The first component is the partial log-likelihood of failure times taking the random effects fixed:

$$l_1 = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \delta_{ij} \left\{ x_{ij}^T \beta + z_{ij}^T b_i - \ln \sum_{t_{kl} \geq t_{ij}} \exp\left(x_{kl}^T \beta + z_{kl}^T b_k \right) \right\}. \tag{9}$$

The second component is related to the distribution associated to the random effects

$$l_2 = -\frac{1}{2} \sum_{g=0}^{d} \left(N \ln 2\pi\theta_g + \sum_{i=1}^{N} \frac{b_{ig}^2}{\theta_g} \right). \tag{10}$$

The algorithm iterates between two steps. First, given an estimate for the variance-covariance matrix $D(\theta)$ one iteration is performed to update the estimate parameters β and b_i. Second, based on the updated values for β and b_i the REML estimator of $D(\theta)$ is used. Once $D(\theta)$ is estimated and updated, the process starts all over again. The details are as follows.

Given values $\beta^{(p)}$ and $b_i^{(p)}$ of the fixed and the random effects, the Newton-Raphson iterative procedure is used for maximizing $l_1 + l_2$ to obtain BLUP estimators $\beta^{(p+1)}$ and $b_i^{(p+1)}$. Let $\eta_{ij} = x_{ij}^T \beta + z_{ij}^T b_i$ and $\eta = (\eta_1^T, \eta_2^T, ..., \eta_N^T)$, where $\eta_i = (\eta_{i1}, \eta_{i2}, ..., \eta_{in_i})^T$. In matrix form, $\eta = X\beta + Zb$, where X and Z are design matrices for the fixed and the random effects, respectively. The random effects are of the form $b = (b_0^T, b_1^T, ..., b_d^T)^T$, where $b_g = (b_{1g}, b_{2g}, ..., b_{Ng})^T$. The Newton-Raphson procedure is carried out as follows:

$$\begin{pmatrix} \beta^{(p+1)} \\ b^{(p+1)} \end{pmatrix} = \begin{pmatrix} \beta^{(p)} \\ b^{(p)} \end{pmatrix} - A^{-1} \begin{pmatrix} 0 \\ \{D(\theta)^{(p)}\}^{-1}\ b^{(p)} \end{pmatrix} + A^{-1} \begin{pmatrix} X^T \\ Z^T \end{pmatrix} \frac{\partial\, l_1}{\partial\, \eta}, \tag{11}$$

where

$$A = \begin{pmatrix} A_{11}\ A_{12} \\ A_{21}\ A_{22} \end{pmatrix} = \begin{pmatrix} X^T \\ Z^T \end{pmatrix} \left[\frac{-\partial^2\, l_1}{\partial\, \eta\ \partial\, \eta^T} \right] \begin{pmatrix} X\ Z \end{pmatrix} + \begin{pmatrix} 0_X & 0_Z \\ 0_Z & \{D(\theta)^{(p)}\}^{-1} \otimes I_N \end{pmatrix},$$

$(p+1)$ and (p) indicate the iterations of the algorithm, $\otimes$ indicates the Kronecker product and I_N the identity matrix of dimension $N \times N$. The dimension of the zero matrices 0_X and 0_Z depends on the dimension of the vectors β and b. Matrix 0_X is a square matrix, while 0_Z has the same number of rows as 0_X, but its number of columns depends on the size of b. The inverse of A will be denoted by M and can be expressed as

$$
\begin{aligned}
M &= \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \\
&= \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1} A_{12} S_{A_{11}}^{-1} A_{21} A_{11}^{-1} & -A_{11}^{-1} A_{12} S_{A_{11}}^{-1} \\ -S_{A_{11}}^{-1} A_{21} A_{11}^{-1} & S_{A_{11}}^{-1} \end{pmatrix},
\end{aligned}
$$

where $S_{A_{11}} = (A_{22} - A_{21} A_{11}^{-1} A_{12})$.

Given the estimated β and b, the REML estimator of θ_g is:

$$
\theta_g^{(p+1)} = \frac{b_g^{(p+1)^T} \; b_g^{(p+1)}}{N - \theta_g^{(p)^{-1}} \; tr(M_{22(g)})}, \tag{12}
$$

where $tr(M_{22(g)})$ indicates the sum of elements of the diagonal of the submatrix of M_{22} related to the component g of the random effects. The algorithm alternates between (11) and (12) to estimate the values of the parameters.

3.1.2 Variance Estimation

The elements needed in the estimation of the variance-covariance matrices for the estimated $\hat{\beta}$ and $\hat{\theta}$ are computed in the iterative procedure. The variance-covariance matrix for the estimate of β is given by M_{11}. The asymptotic variance of $\hat{\theta}_g$ is given by:

$$
2\hat{\theta}_g^2 \left\{ N - 2\hat{\theta}_g^{-1} \; tr(M_{22(g)}) + \hat{\theta}_g^{-2} \; tr(M_{22(g)}^2) \right\}^{-1}. \tag{13}
$$

3.2 Approximate Marginal Likelihood Method

Using the derivation of a penalized likelihood solution obtained by Breslow and Clayton (1993) for the generalized linear mixed model assuming Gaussian random effects, Ripatti and Palmgren (2000) presented a parallel approximation for model (3).

3.2.1 Penalized Partial Likelihood

Ripatti and Palmgren (2000) approximate the marginal likelihood (6) using the Laplace approximation. For Gaussian random effects the marginal likelihood (6) can be rewritten as:

$$L^M(\beta, \lambda_0, b) = c|D(\theta)|^{-\frac{N}{2}} \int e^{-\kappa(b)} db,$$

where

$$\kappa(b) = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \{\delta_{ij}(\ln \lambda_0(t_{ij}) + x_{ij}^T\beta + z_{ij}^T b_i) - \Lambda_0(t_{ij}) \exp(x_{ij}^T\beta + z_{ij}^T b_i)\} - \frac{1}{2} b_i^T \{D(\theta)\}^{-1} b_i \right]. \quad (14)$$

Let κ', κ'' denote the first and the second order partial derivatives of κ with respect to b. Ignoring a multiplicative constant, the approximation of the logarithm of the marginal likelihood takes the form:

$$l^M(\beta, \lambda_0, b) \approx -\frac{N}{2} |\ln D(\theta)| - \frac{1}{2} \ln |\kappa''(\tilde{b})| - \kappa(\tilde{b}), \quad (15)$$

with $\tilde{b} = \tilde{b}(\beta, \theta)$ the solution to $\kappa'(\tilde{b}) = 0$.

Ripatti and Palmgren (2000) show that, for fixed θ, the values $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$, which maximize the penalized log-likelihood (14), also maximize the penalized partial log-likelihood

$$l^{PPL}(\beta, \lambda_0, b) = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \delta_{ij} \left\{ (x_{ij}^T\beta + z_{ij}^T b_i) - \ln \sum_{t_{kl} \geq t_{ij}} \exp\left(x_{kl}^T\beta + z_{kl}^T b_k\right) \right\} - \frac{1}{2} b_i^T D(\theta)^{-1} b_i \right]. \quad (16)$$

Note that the penalized partial log-likelihood is just the sum of the elements on equations (9) and (10) containing the parameters of interest (β and b). The estimating equations for $\beta(\theta)$ and $b(\theta)$, for a given θ, are of the form:

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i}\delta_{ij}\left\{x_{ij}-\frac{x_{ij}\exp\left(x_{ij}^T\beta+z_{ij}^Tb_i\right)}{\sum_{t_{kl}\geq t_{ij}}\exp\left(x_{kl}^T\beta+z_{kl}^Tb_k\right)}\right\}=0, \qquad (17)$$

$$\sum_{i=1}^{N}\left[\sum_{j=1}^{n_i}\delta_{ij}\left\{z_{ij}-\frac{z_{ij}\exp\left(x_{ij}^T\beta+z_{ij}^Tb_i\right)}{\sum_{t_{kl}\geq t_{ij}}\exp\left(x_{kl}^T\beta+z_{kl}^Tb_k\right)}\right\}-\{D(\theta)\}^{-1}b_i\right]=0, \qquad (18)$$

where the product of a scalar by a vector results in multiplying each elements of the vector by the scalar.

Ripatti and Palmgren (2000) propose to find $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$ by alternating between solving the equations (17) and (18). Once $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$ are computed, θ is updated by maximizing the approximate profile likelihood derived from (15):

$$l^M(\beta,\lambda_0,\theta)\approx-\frac{N}{2}|\ln D(\theta)|-\frac{1}{2}\ln|\kappa''(\hat{b})|-\frac{1}{2}\hat{b}^T\{D(\theta)\}^{-1}\hat{b}. \qquad (19)$$

Ripatti and Palmgren (2000) propose to use $\kappa''_{PPL}(b)=(\partial^2 l^{PPL})/(\partial b\partial b^T)$ instead of $\kappa''(b)$, given its better empirical performance. An estimating equation for θ can be obtained after differentiation of (19) and some simplifications. In the particular case of a diagonal $D(\theta)$ the solution of the estimating equation takes the following simple form:

$$\hat{\theta}_g=\frac{\hat{b}_g^T\ \hat{b}_g+\mathrm{tr}\{\kappa''_{PPL}(\hat{b})^{-1}_{(g)}\}}{N}. \qquad (20)$$

where $tr\{\kappa''_{PPL}(\hat{b})^{-1}_{(g)}\}$ indicates the sum of the elements of the diagonal of the submatrix of $\kappa''_{PPL}(\hat{b})^{-1}$ associated with the g component of the random effects.

3.2.2 Variance Estimation

Estimates of the variance-covariance matrix of the fixed effects can be obtained using standard Cox regression with the estimated random effects as an offset. In order to estimate the variance-covariance matrix of $\hat{\theta}$ it is necessary to differentiate (19) twice with respect to θ and take the expectation with respect to b. Under the assumed diagonal form of $D(\theta)$, the variance for $\hat{\theta}$ is given by

$$\text{var}(\hat{\theta}_g) = 2\ \hat{\theta}_g^2 \left[N + \frac{1}{\hat{\theta}_g^2} tr\left\{ \kappa''_{PPL}(\hat{b})^{-1}_{(g)} \kappa''_{PPL}(\hat{b})^{-1}_{(g)} \right\} - \frac{2}{\hat{\theta}_g} tr\left\{ \kappa''_{PPL}(\hat{b})^{-1}_{(g)} \right\} \right]^{-1} \quad (21)$$

3.3 Bayesian Estimation Approach

Ducrocq and Casella (1996) have proposed a Bayesian approach to estimate the parameters of the distribution of the random effects. In this approach the variance components related to the distribution of the random effects are estimated from their marginal posterior distribution after integrating out β and b. As this integration can not be performed analytically, the Laplace approximation is used.

3.3.1 Laplace approximation of the marginal posterior distribution

Applying the Bayes theorem, the joint posterior density for model (3) is proportional to

$$L^B(\beta, b, \theta \mid y) \propto L(y \mid \beta, b) \times \pi_0(b \mid \theta) \times \pi_0(\beta) \times \pi_0(\theta). \quad (22)$$

In this expression, the first factor is the partial likelihood (see (9)), while $\pi_0(b \mid \theta)$ is the joint normal density (see (10)).

Ducrocq and Casella (1996) assume a flat prior for θ and β

$$\pi_0(\theta) \propto 1 \text{ and } \pi_0(\beta) \propto 1.$$

Therefore, the log joint posterior density is given by the sum of equations (9) and (10). It is interesting to note that the term "posterior density" is in fact used here for convenience, acknowledging that it is obtained using the partial likelihood and not the full likelihood.

According to the Bayesian principle, estimation of the vector of variance components θ of the random effects should be based on its marginal posterior distribution after integrating out the nuisance parameters β and b:

$$L^P(\theta \mid y) = \int L^B(\beta, b, \theta \mid y) d\beta db \quad (23)$$

As this integration cannot be performed analytically, Ducrocq and Casella

(1996) propose to approximate this marginal posterior density using the Laplace approximation. More precisely, for any given value θ^* of θ, they show that

$$L^P(\theta^* \mid y) \approx \int \exp\left\{ l^B(\hat{\Psi}_{\theta^*} \mid y, \theta^*) - \frac{1}{2}(\Psi - \hat{\Psi}_{\theta^*})^T H_{\theta^*}(\Psi - \hat{\Psi}_{\theta^*}) \right\} \beta db, \quad (24)$$

where H_{θ^*} is the negative Hessian matrix of the joint posterior distribution computed at the maximum $\hat{\Psi}_{\theta^*} = (\hat{\beta}_{\theta^*}, \hat{b}_{\theta^*})$ of $l^B(\beta, b \mid y, \theta = \theta^*)$. Recognizing under the integral sign of this last equation the kernel of a multivariate normal density with mean $\hat{\Psi}_{\theta^*}$ and variance H_{θ^*}, Ducrocq and Casella derive the following approximation of the marginal posterior density for any θ^* of θ:

$$l^P(\theta^* \mid y) \approx \text{constant} + l^B(\hat{\Psi}_{\theta^*} \mid y, \theta^*) - \frac{1}{2} \ln \mid H_{\theta^*} \mid . \quad (25)$$

For any fixed value θ^* of θ, the log joint posterior density is maximized using a limited memory quasi-Newton method (Liu and Nocedal, 1989) to obtain point estimates $\hat{\beta}_{\theta^*}$ and $\hat{b}_{\theta^*}$ of β and b. The negative Hessian matrix H_{θ^*} is then computed at this maximum. Based on $\hat{\Psi}_{\theta^*}$ and H_{θ^*}, the approximate log marginal posterior density at θ^*, obtained from formula (25), is computed.

The method of the simplex (Nelder and Mead, 1965) is then used, in a upper level of iterations, to select the value of θ which maximizes this approximate log marginal posterior distribution. The mode of this approximate log marginal posterior distribution is taken as the point estimate for θ. Note that one could also use another point estimates, given that we have the whole distribution of θ.

3.3.2 Variance Estimation

As for the Ripatti and Palmgren (2000) approach, estimates of the variance of the fixed effects β are easily obtained using standard Cox regression with the estimated random effects as an offset.

Estimates of the standard error of $\hat{\theta}$, as well as other point estimates of the distribution of $\hat{\theta}$, can be derived from the knowledge of the full marginal posterior density. To avoid repeated computations of (25), and in particular of the negative Hessian matrix H for many different values of θ, Ducrocq and Casella (1996) propose to summarize the general characteristics of the distribution (25) through the computation of its first three moments by numerical integration based on the Gauss-Hermite quadrature. It is worth mentioning that, in general, this distribution appears to be substantially skewed. It follows that computing the standard deviation of the marginal posterior distribution

of θ and using it as standard error of the parameter can lead to overestimation. It is important to note that if we are interested to test whether $\theta > 0$, we can use the whole distribution to compute the confidence interval, without relying on asymptotic theory, what can be seen as an advantage.

3.4 The EM Algorithm with the Laplace Approximation

Cortiñas and Burzykowski (2004) developed an estimation method based on the use of the Laplace approximation at the E-step in the EM algorithm.

3.4.1 The E-step

In the E-step the expectation of the logarithm of the likelihood (8), conditional on the observed data and on the current values $\beta^{(p)}$, $\theta^{(p)}$ and $\lambda_0^{(p)}$ of parameters β, θ and λ_0, respectively, is computed. The expectation, denoted by $Q(\beta, \theta, \lambda_0)$, can be written as:

$$Q(\beta, \theta, \lambda_0) = Q_1(\beta, \lambda_0) + Q_2(\theta), \tag{26}$$

where

$$\begin{aligned} Q_1(\beta, \lambda_0) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \Big[& \delta_{ij} \left\{ \ln \lambda_0(t_{ij}) + x_{ij}^T \beta + z_{ij}^T \mathrm{E}(b_i) \right\} \\ & - \Lambda_0(t_{ij}) \exp \left\{ x_{ij}^T \beta + \ln \mathrm{E}(e^{z_{ij}^T b_i}) \right\} \Big] \end{aligned} \tag{27}$$

and

$$Q_2(\theta) = -\frac{1}{2} \sum_{g=1}^{d} \left\{ N \ \ln(2\pi\theta_g) + \sum_{i=1}^{N} \frac{\mathrm{E}(b_{ig}^2)}{\theta_g} \right\}, \tag{28}$$

with E(.) denoting the expected values. To simplify the notation, the dependence of the expected values in (27) and (28) on the observed data and $\beta^{(p)}$, $\theta^{(p)}$ and $\lambda_0^{(p)}$ has been suppressed. Note that $Q_1(\beta, \lambda_0)$ is just the conditional log-likelihood (4), where the random effects b_i are replaced by their expectations. It is important to remark, that the expectations in equation (27) and (28) will not be available in a closed-form. The conditional expectations that need to be computed involve integrals of the form

$$\mathrm{E}\{g(b_i)\} = \frac{\int g(b_i) e^{l_i^C\left(\beta^{(p)},\lambda_0^{(p)},b_i\right)+\ln f(b_i,\theta^{(p)})} db_i}{\int e^{l_i^C(\beta^{(p)},\lambda_0^{(p)},b_i)+\ln f(b_i,\theta^{(p)})} db_i} \ . \tag{29}$$

Cortiñas and Burzykowski (2004) propose to use the Laplace formula to compute these expectations. Using the formula results in the approximations

$$\mathrm{E}\{g(b_i)\} \approx g(\widetilde{b}_i) \ , \tag{30}$$

where $\widetilde{b}_i$ is an isolated global minimum of

$$k(b_i) = -\frac{1}{n_i}\left\{l_i^C\left(\beta^{(p)}, \lambda_0^{(p)}, b_i\right) + \ln f(b_i, \theta^{(p)})\right\} \ . \tag{31}$$

The set of initial values for β and λ_0 are obtained using the Cox regression without random effects. The initial values for θ can be specified by taking $D(\theta)$ equal to, e.g., the identity matrix.

3.4.2 The M-step

In the M-step new estimates $\beta^{(p+1)}$ and $\theta^{(p+1)}$ are found by maximizing the functions Q_1 and Q_2, respectively. To estimate β the profile likelihood approach is used. Assuming no ties, in order to keep notation simple, the value of the baseline hazard which maximizes Q_1 is

$$\lambda_m^{(p+1)} = \frac{1}{\sum_{t_{kl}\geq t_m} \exp\{x_{kl}^T\beta^{(p)} + e^{z_{kl}^T b_k}\}}, \tag{32}$$

where $\lambda_m = \lambda_0(t_m)$ and t_m $(m = 1, \ldots, r)$ are the distinct uncensored failure times. Substituting (32) into Q_1 gives the following profile-likelihood for β:

$$Q_1'(\beta) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\delta_{ij}\left[x_{ij}^T\beta - \ln\sum_{t_{kl}\geq t_{ij}} \exp\left\{x_{kl}^T\beta + \ln \mathrm{E}(e^{z_{kl}^T b_k})\right\}\right]. \tag{33}$$

The form of (33) resembles that of the partial log-likelihood for the Cox proportional hazards model with offsets $\ln \mathrm{E}(e^{z_{ij}^T b_i})$. New value $\beta^{(p+1)}$ of β is obtained by maximizing Q_1' using standard software for the Cox model, as in the method proposed by Ripatti and Palmgren (2000).

Given that the density of the random effects b_i belongs to the exponential family, the estimation of $D(\theta)$ is generally straightforward. Hence, maximizing Q_2 leads to the estimator

$$\hat{D}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left(b_i b_i^T\right). \tag{34}$$

3.4.3 Variance Estimation

The variance-covariance matrix of the solution $(\hat{\beta}, \hat{\lambda}_0, \hat{\theta})$ obtained from the EM algorithm, can be estimated using the inverse of the observed information matrix computed from the formula proposed by Louis (1982):

$$\mathrm{I}(\beta, \lambda_0, \theta) = \left[\mathrm{E}\left\{-l^{A''}(\beta, \lambda_0, \theta)\right\} - \mathrm{E}\left\{l^{A'}(\beta, \lambda_0, \theta) l^{A'}(\beta, \lambda_0, \theta)^T\right\}\right], \tag{35}$$

where $l^{A'}$ and $l^{A''}$ are the first and the second derivatives with respect to $(\beta, \lambda_0, \theta)$ of the logarithm of the "augmented" likelihood (8).

In order to compute standard error for the fixed parameters and the variance component of the random effects, it would be necessary to invert $I(\beta, \lambda_0, \theta)$. The dimension of the matrix $I(\beta, \lambda_0, \theta)$ can be very large, since it depends on λ_0 and hence on the number of distinct uncensored failure times. Cortiñas and Burzykowski (2004) proposed to estimate the standard error of the parameter of interest by inverting only the relevant blocks of the matrix, corresponding to β and θ.

4 Simulation Study

A simulation study was carried out to compare the performance of the methods of McGilchrist and Aisbett (1991), Ducrocq and Casella (1996), Ripatti and Palmgren (2000) and Cortiñas and Burzykowski (2004). The data were generated using the following proportional hazards model:

$$\lambda(t_{ij}|\beta, b_{i0}, b_{i1}) = \lambda_0(t_{ij}) e^{b_{i0} + x_{ij}^T(\beta + b_{i1})}, \tag{36}$$

with

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N_2\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ & \sigma_1^2 \end{pmatrix}\right\}. \tag{37}$$

Model (36) corresponds to the setting of a multi-center clinical trial, in which heterogeneity appears both in the center-specific baseline hazards, as well as associated to the covariate. Parameters of the model were chosen to mimic data available in a real bladder cancer clinical trial database (Royston, Parmar and Sylvester, 2004). In this simulation study we considered the same distribution of patients over centers as in the real dataset, namely 2323 patients accrued by 37 centers. Distribution of patients over centers was as follows: 21, 23, 23, 25, 26, 30, 30, 32, 34, 34, 34, 35, 35, 35, 37, 39, 41, 42, 42, 43, 52, 52, 53, 56, 61, 63, 66, 72, 85, 86, 91, 104, 116, 120, 155, 183, 247.

Two simulation settings were considered. The first simulation setting assumed moderately censored data (around 40%), while the second one assumed highly censored data (around 60%). In both settings we simulated data using different combinations of values of σ_0^2 and σ_1^2, namely 0.04/0.08, 0.08/0.04, 0.08/0.08 and 0.4/0.8. A constant baseline hazard of 0.077 was used. In order to have a better idea of the interpretation of the values of σ_0^2, σ_1^2 and λ_0, one might consider the spread of the median time-to-event from center to center. In our simulation b_{i0} and b_{i1} are normally distributed, thus the resulting distribution function for the median time to event is log-normal with parameters $\ln(\ln 2) - \ln \lambda_0 - \beta x_{ij}$ and $\sigma_0^2 + \sigma_1^2 x_{ij}^2$. For patients with $x_{ij} = 0$, the distribution does not depend neither on β, nor on σ_1^2. For such patients, Table 1 presents, for each particular value of σ_0^2, the interval containing the median time-to-event of 90% of the centers.

Table 1
Interpretation of σ_0^2.

σ_0^2	Median time-to-event for patients with $x_{ij} = 0$
0.04	(6.5 yrs - 12.5 yrs)
0.08	(5.7 yrs - 14.4 yrs)
0.4	(3.2 yrs - 25.5 yrs)

In the next sections we describe details of both settings.

4.1 Moderate Censoring Setting

In this setting we consider a covariate x_{ij}, which divides the population in two groups: 30% of the patients have $x_{ij} = 0$ and 70% have $x_{ij} = 1$. The covariate can be seen as corresponding to, e.g., a prognostic index. The parameter β was set equal to 0.7, which corresponds to the estimated value for the real dataset. Given β and λ_0, we can compute the median time-to-event for a model without the random effects, which equal 9 in the group defined by $x_{ij} = 0$ and 4.5 in the group defined by $x_{ij} = 1$. Table 2 presents, for each particular value of σ_0^2

and σ_1^2, the intervals containing the hazard ratio of the effect of x_{ij} for 90% of the centers. Also, a corresponding interval for the median time-to-event in the group of patients with $x_{ij} = 1$ is given.

For each parameter setting, 250 datasets were generated in the following way. First, $N = 37$ random effects for the overall center effect and $N = 37$ random effects for the center-specific covariate effect were generated according to (37). We considered an accrual period (AP) of 1065 days (about 3 years) and a further follow up period (FP) of 2440 days (about 6.7 yrs). The actual observation for a patient was the minimum of the time to event and the time at risk. The former was generated using an exponential random variable with parameter $\lambda(t_{ij}|\beta_i, b_{i0}, b_{i1})$ given by (36). The time at risk for a patient who entered in the study as k-th subject at time $\frac{kAP}{2323}$ was defined as

$$\frac{AP(2323-k)}{2323} + FP.$$

Table 2
Interpretation of σ_1^2.

σ_0^2	σ_1^2	Hazard ratio	Median time-to-event for patients with $x_{ij} = 1$
0.04	0.08	(1.26 - 3.21)	(2.6 yrs - 7.9 yrs)
0.08	0.04	(1.45 - 2.80)	(2.6 yrs - 7.9 yrs)
0.08	0.08	(1.26 - 3.21)	(2.4 yrs - 8.7 yrs)
0.4	0.8	(0.46 - 8.77)	(0.8 yrs - 27.1 yrs)

These particular choices of the parameters resulted in approximately 60% of the individuals experiencing the event of interest.

4.2 Heavy Censoring Setting

In this setting, it was assumed that x_{ij} represented the treatment assignment. Hence, an equal split of patients in the two treatment groups ($x_{ij} = 0$ and $x_{ij} = 1$) was used. Assuming that the clinical trial takes place in good prognosis bladder cancer patients, and that the experimental treatment leads to 20% increase in the median disease free interval, i.e., from 9 to 10.8 years, we used a baseline hazard of 0.077 as in the "moderate censoring" setting, with β equal to -0.182. Given the values of β, λ_0, σ_0^2 and σ_1^2, we can compute intervals containing the hazard ratio of the effect of the covariate and median time-to-event for 90 % of the centers (Table 3).

For each parameter setting, 250 datasets were generated. In this setting we considered an accrual period of 621 days (about 1.7 yrs) and a further follow

Table 3
Interpretation of σ_1^2.

σ_0^2	σ_1^2	Hazard ratio	Median time-to-event for patients with $x_{ij} = 1$
0.04	0.08	(0.52 - 1.33)	(6.1 yrs - 19.1 yrs)
0.08	0.04	(0.60 - 1.16)	(6.1 yrs - 19.1 yrs)
0.08	0.08	(0.52 - 1.33)	(5.6 yrs - 20.8 yrs)
0.4	0.8	(0.19 - 3.63)	(1.8 yrs - 65.5 yrs)

up period of 2192 days (about 6 years). The generating mechanism for the data was similar to the one used in "moderate censoring" setting. As a result of the choices of the parameters, approximately 40% of the individuals in the datasets experienced the event of interest.

McGilchrist and Aisbett's approach was implemented using SAS-IML v8.2. The iterative procedure was stopped if the maximum of the relative difference between the fixed effects and variance estimates for two consecutive iterations was smaller than 10^{-3}. The method proposed by Ducrocq and Casella (1996) was implemented with The Survival Kit (Ducrocq and Sölkner , 1994; Ducrocq and Sölkner, 1998)(www.boku.ac.at/nuwi/software/softskit.htm), a package of Fortran programs developed in the field of animal genetics to estimate survival models with random effects. The joint estimation of two variance components was not implemented in the original version. Therefore, we used a modified version proposed by Legrand et al. (2004). In this case, the convergence criterion required that the standardized norm of the vector of first derivatives of $l^B(b, b|y, \theta = \theta^*)$ at its maximum had to be less than 10^{-8}. The EM algorithm proposed by Cortiñas and Burzykowski (2004) was implemented using SAS-IML v8.2. The EM algorithm stopped when the maximum of the absolute changes for the fixed effects, the variance estimates and the loglikelihood was smaller than 10^{-5}. The method proposed by Ripatti and Palmgren (2000) was applied using the S+ functions developed by Therneau (2003). In this case the convergence criterion was the relative change in log likelihood smaller than 10^{-4}.

5 Results of the Simulations

5.1 Moderate Censoring Setting

Table 4 presents the results of the simulation for the moderate censoring setting. In this setting, none of the methods used in the simulations experienced convergence difficulties. The parameter β was in general estimated well by all

the methods, with a relative absolute bias less then 4% in any of the considered cases. The bias increased with increasing σ_0^2 and σ_1^2. The bias of the estimates obtained by the Ripatti and Palmgren (2000) approach, in the case of $\sigma_0^2 = 0.04$ and $\sigma_1^2 = 0.08$, was higher than for the other methods. On the other hand, the approach of Ducrocq and Casella (1996) produced the largest bias when $\sigma_0^2 = 0.4$ and $\sigma_1^2 = 0.8$. The variability of estimates of β, measured by the empirical standard error, was similar for all the methods compared. In general, the Ripatti and Palmgren approach produced fixed-effects estimates with the largest empirical standard error. It is important to note, that the model-based estimates produced by Ducrocq and Casella (1996) approach were closer to the empirical standard error than for the other methods.

Table 4
Moderate censoring setting. The mean estimates for 250 simulated datasets for the 4 different methods. In parentheses: the mean model-based and empirical (first and second number) standard error.

Method	$\hat{\beta}$	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$
	$\sigma_0^2 = 0.04$ and $\sigma_1^2 = 0.08$		
McGilchrist	0.705(0.074;0.073)	0.040(0.169;0.023)	0.078(0.396;0.034)
Ripatti	0.715(0.119;0.094)	0.038(0.023;0.024)	0.079(0.026;0.039)
EM-Laplace	0.702(0.055;0.079)	0.044(0.021;0.023)	0.083(0.030;0.031)
Ducrocq	0.703(0.081;0.078)	0.042(0.033;0.025)	0.079(0.047;0.039)
	$\sigma_0^2 = 0.08$ and $\sigma_1^2 = 0.04$		
McGilchrist	0.701(0.067;0.068)	0.079(0.444;0.027)	0.043(0.161;0.028)
Ripatti	0.702(0.083;0.095)	0.076(0.025;0.030)	0.039(0.021;0.030)
EM-Laplace	0.693(0.055;0.071)	0.075(0.022;0.028)	0.055(0.024;0.026)
Ducrocq	0.713(0.073;0.079)	0.085(0.041;0.039)	0.039(0.039;0.031)
	$\sigma_0^2 = 0.08$ and $\sigma_1^2 = 0.08$		
McGilchrist	0.703(0.074;0.079)	0.084(0.459;0.032)	0.075(0.343;0.038)
Ripatti	0.702(0.082;0.085)	0.077(0.025;0.036)	0.077(0.026;0.042)
EM-Laplace	0.696(0.055;0.085)	0.076(0.028;0.031)	0.080(0.029;0.032)
Ducrocq	0.706(0.081;0.082)	0.084(0.046;0.039)	0.082(0.053;0.042)
	$\sigma_0^2 = 0.4$ and $\sigma_1^2 = 0.8$		
McGilchrist	0.689(0.160;0.150)	0.405(2.760;0.121)	0.797(5.808;0.213)
Ripatti	0.691(0.122;0.165)	0.383(0.101;0.121)	0.770(0.195;0.213)
EM-Laplace	0.698(0.156;0.166)	0.375(0.083;0.092)	0.765(0.189;0.219)
Ducrocq	0.672(0.161;0.167)	0.402(0.162;0.140)	0.752(0.218;0.194)

The estimates σ_0^2 and σ_1^2 for all the methods were on average comparable. The version of the EM algorithm proposed by Cortiñas and Burzykowski (2004) yielded estimates with, in general, the smallest empirical variability, while

Ducrocq and Casella's approach gave estimates with the largest variability. The model-based standard errors for the approach of McGilchrist and Aisbett severely overestimated the true variability. The Ducrocq and Casella approach tended to overestimate the empirical variability, while the Ripatti and Palmgren method and the version of the EM algorithm proposed by Cortiñas and Burzykowski (2004) tended to underestimate it. The model-based standard errors produced by the version of the EM algorithm proposed by Cortiñas and Burzykowski (2004) were in most of the cases closer to the empirical standard error than for the other methods. Table 5 shows the mean squared error for each parameter in each of the settings studied. In the setting with small variances for the random effects, the mean squared errors (MSE) are comparable. For the case of $\sigma_0^2 = 0.4$ and $\sigma_1^2 = 0.8$, larger values of the mean squared errors were found, but no clear pattern could be seen that would indicate that one method is preferable to all others in all circumstances.

Table 5
Mean Squared Error for the parameters for the 4 different methods ($\times 10^{-3}$).

	Moderate censoring setting			Heavy censoring setting		
Method	$\hat{\beta}$	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$	$\hat{\beta}$	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$
	$\sigma_0^2 = 0.04$ and $\sigma_1^2 = 0.08$					
McGilchrist	5.35	0.53	1.16	6.97	0.53	2.81
Ripatti	9.06	0.58	1.52	17.16	0.50	2.12
EM-Laplace	6.25	0.55	0.97	9.03	0.49	0.90
Ducrocq	6.09	0.63	1.52	6.95	0.58	2.30
	$\sigma_0^2 = 0.08$ and $\sigma_1^2 = 0.04$					
McGilchrist	4.63	0.73	0.79	7.06	1.16	1.45
Ripatti	9.03	0.92	0.90	26.28	0.84	1.30
EM-Laplace	5.09	0.81	0.90	5.97	0.63	0.76
Ducrocq	6.41	1.55	0.96	6.21	1.23	1.45
	$\sigma_0^2 = 0.08$ and $\sigma_1^2 = 0.08$					
McGilchrist	6.25	1.04	1.47	7.57	1.23	2.11
Ripatti	7.23	1.31	1.77	7.97	1.03	2.61
EM-Laplace	7.24	0.98	1.02	8.48	0.97	0.97
Ducrocq	6.76	1.54	1.77	7.06	1.17	2.75
	$\sigma_0^2 = 0.4$ and $\sigma_1^2 = 0.8$					
McGilchrist	22.62	14.67	45.38	32.83	13.98	64.44
Ripatti	27.31	14.93	46.27	26.28	13.99	64.73
EM-Laplace	27.56	9.09	49.19	25.07	9.06	41.16
Ducrocq	28.67	19.60	39.94	24.82	14.68	17.73

5.2 Heavy Censoring Setting

Table 6 shows the results of the simulation for the heavy censoring setting. In this setting, Ripatti and Palmgren's approach and the EM algorithm with the Laplace approximation experienced convergence problems (up to 14 and 17 % respectively). The mean estimated values reported here were based only on the cases when convergence was reached.

Table 6

Heavy censoring setting. The mean estimates for 250 simulated datasets for the 4 different methods. In parentheses: the mean model-based and empirical (first and second number) standard error.

Method	$\hat{\beta}$	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$	Convergence(%)
	$\sigma_0^2 = 0.04$ and $\sigma_1^2 = 0.08$			
McGilchrist	-0.173(0.085;0.083)	0.039(0.149;0.023)	0.081(0.309;0.053)	100
Ripatti	-0.166(0.092;0.130)	0.044(0.021;0.022)	0.080(0.036;0.046)	86.4
EM-Laplace	-0.168(0.067;0.094)	0.043(0.015;0.022)	0.079(0.026;0.030)	83.2
Ducrocq	-0.167(0.084;0.082)	0.042(0.031;0.024)	0.080(0.031;0.048)	100
	$\sigma_0^2 = 0.08$ and $\sigma_1^2 = 0.04$			
McGilchrist	-0.181(0.077;0.084)	0.080(0.411;0.034)	0.043(0.124;0.038)	100
Ripatti	-0.176(0.100;0.162)	0.079(0.024;0.029)	0.039(0.034;0.036)	94.4
EM-Laplace	-0.176(0.067;0.077)	0.078(0.022;0.026)	0.056(0.015;0.020)	88.6
Ducrocq	-0.171(0.077;0.078)	0.082(0.042;0.035)	0.042(0.056;0.038)	100
	$\sigma_0^2 = 0.08$ and $\sigma_1^2 = 0.08$			
McGilchrist	-0.156(0.083;0.083)	0.081(0.405;0.035)	0.071(0.243;0.045)	100
Ripatti	-0.167(0.086;0.088)	0.081(0.024;0.032)	0.082(0.037;0.051)	97.2
EM-Laplace	-0.168(0.066;0.091)	0.082(0.028;0.031)	0.082(0.029;0.031)	94.8
Ducrocq	-0.169(0.086;0.834)	0.083(0.044;0.034)	0.087(0.072;0.052)	100
	$\sigma_0^2 = 0.4$ and $\sigma_1^2 = 0.8$			
McGilchrist	-0.139(0.161;0.176)	0.383(2.615;0.117)	0.730(4.809;0.244)	100
Ripatti	-0.156(0.143;0.160)	0.392(0.078;0.118)	0.765(0.180;0.252)	100
EM-Laplace	-0.155(0.129;0.156)	0.385(0.085;0.094)	0.766(0.165;0.200)	100
Ducrocq	-0.160(0.165;0.156)	0.406(0.151;0.121)	0.767(0.227;0.129)	100

Fixed effects β were in general well estimated for $\sigma_0^2 = 0.04$ and $\sigma_1^2 = 0.08$ and for $\sigma_0^2 = 0.08$ and $\sigma_1^2 = 0.04$. For these cases, the relative bias was smaller than 9 %. The picture was somewhat different when the variances increased. The relative bias for the fixed effects reached 24% (McGilchrist and Aisbett's approach), but in general it was smaller than 15%. Similar to the previous "moderate censoring" setting, the estimates for the Ripatti and Palmgren approach showed, in general, the largest empirical variability, while those for the Ducrocq and Casella method produced, in general, estimates with the smallest variability. Note also, that the model-based standard error for the approach proposed by the Ducrocq and Casella (1996) was the closest to the empirical value.

The estimates of variances of the random effects, similarly to the previous setting, were comparable for all the methods. The McGilchrist and Aisbett

method produced heavily biased model-based standard errors. The Ripatti and Palmgren approach and the version of the EM algorithm proposed by Cortiñas and Burzykowski (2004) underestimated the empirical variability, while the method of Ducrocq and Casella overestimated it. The version of the EM algorithm proposed by Cortiñas and Burzykowski (2004), in most of the cases, yielded model-based standard errors that were closer to the empirical standard error than for the other methods.

In terms of mean squared errors, similar conclusion to the one obtained in the "moderate censoring" setting can be drawn. Note, that when one of the variance of the random effects was considered equal to 0.04, the method proposed by Ripatti and Palmgren (2000) produced higher values of MSE for the fixed effect than the rest of the estimation methods. In general, no clear pattern can be observed which allows to select the best estimation procedure.

6 Concluding Remarks

Proportional hazards models with multivariate random effects offer several advantages over univariate shared frailty models (Xue and Brookmeyer, 1996), especially when survival times from the same cluster are negatively associated. The main stumbling block in the use of the former models are estimation methods. In this paper the performance of four estimation methods was compared.

The results clearly show problems with the computation of the standard error of the estimated variance components for McGilchrist and Aisbett's approach. In terms of the point estimates, the four methods were in general comparable. However, in the heavy censoring setting, when the variance of the random effects was large (0.4 and 0.8), McGilchrist and Aisbett's approach showed larger bias. It is also important to mention that in that setting Ripatti and Palmgren's approach, as well as the method proposed by Cortiñas and Burzykowski (2004), experienced convergence problems. The non-convergence rate for the method proposed by Cortiñas and Burzykowski (2004) was somewhat higher. Ducrocq and Casella's approach produced conservative standard errors of the estimated variance component, while the proposed version of the EM algorithm by Cortiñas and Burzykowski (2004) and the Ripatti and Palmgren method tended to underestimate the true standard error. It can also be noted that in general the distribution of θ appear to be substantially skewed. For this reason, it is worth noting that Ducrocq and Casella's approach can provide the whole distribution of the parameter θ, which can be usefull for testing purposes.

Taking into account the results obtained in the simulations study, we can conclude that the McGilchrist and Aisbett approach suffers from serious problems

in the estimation of the standard errors of the variance components, and for this reason it should rather not be used. The other three methods produce comparable point estimates. The method proposed by Ducrocq and Casella (1996) yields conservative estimates of the standard errors of the variance components and does not suffer from convergence problems, what may be seen as an advantage. On the other hand the estimates obtained by the modified version of the EM algorithm proposed by Cortiñas and Burzykowski (2004) seem to express the smallest empirical variability. Due to these differences, more investigation is needed before a definitive choice between the methods of Ducrocq and Casella (1996), Ripatti and Palmgren (2000) and Cortiñas and Burzykowski (2004) can be made.

Acknowledgment

The authors gratefully acknowledge support from FWO-Vlaanderen Research Project "Sensitivity Analysis for Incomplete and Coarse Data" and Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

Breslow, N.E. and Clayton, D.G., 1993. Approximate inference in generalized linear models. Journal of the American Statistical Association 88, 9–25.

Clayton, D.G.,1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65, 141–151.

Cortiñas Abrahantes, J. and Burzykowski, T., 2004. A version of the EM algorithm for proportional hazards model with random effects. Biometrical Journal (to appear). 2004.

Ducrocq, V. and Sölkner, J., 1994. The Survival Kit, a FORTRAN package for the analysis of survival data. 5th World Cong. Genet. Appl. Livest. Prod. 22, 51–52. Dep. Anim. Poultry Sci., Univ. of Guelph, Guelph, Ontario, Canada.

Ducrocq, V. and Casella, G., 1996. A Bayesian analysis of mixed survival models. Genet. Sel. Evol. 28, 505–529.

Ducrocq, V. and Sölkner, J., 1998. The Survival Kit – V3.0, a package for large analyses of survival data. 6th World Cong. Genet. Appl. Livest. Prod. 27, 447–448. Anim. Genetics and Breeding Unit, Univ. of New England, Armidale, Australia.

Hougaard, P., 2000. Analysis of Multivariate Survival Data. New York: Springer-Verlag.

Legrand, C., Janssen, P., Duchateau, L., Ducrocq, V. and Sylvester, R., 2004.

A Bayesian approach to the estimation of proportional hazards model with random effects. In preparation.

Liu, D. C. and Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization. Mathematical Programming 45, 503–528.

Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society (Series B) 44(2), 190–200.

McGilchrist, C.A. and Aisbett, C.W., 1991. Regression with frailty in survival analysis. Biometricsf 47, 461–466.

McGilchrist, C.A., 1993. REML estimation for survival models with frailty. Biometrics 49, 221–225.

McGilchrist, C.A., 1994. Estimation in generalized mixed models. Journal of the Royal Statistical Society (Series B) 56(1), 61–69.

Nelder, J.A. and Mead, R., 1965. A Simplex method for function minimization. Computer Journal 7, 308–313.

Oakes, D., 1982. A model for association in bivariate survival data. Journal of the Royal Statistical Society (Series B) 44, 414-422.

Ripatti, S. and Palmgren, J., 2000. Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 56, 1016–1022.

Ripatti, S., Larsen, K. and Palmgren, J., 2002. Maximum Likelihood Inference for Multivariate Frailty Models Using an Automated Monte Carlo EM Algorithm. Lifetime Data Analysis 8, 349–360.

Royston, P., Parmar, M.K.B. and Sylvester, R., 2004. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Statistics in Medicine 23, 907–926.

Schall, R., 1991. Estimation in generalised linear models with random effects. Biometrika 78, 719–727.

Therneau, T., 2003. On mixed effect Cox models, sparse matrices, and modelling data from large pedigree. Technical report july 2003.

Vaida F. and Xu, R., 2000. Proportional hazards model with random effects. Statistics in Medicine 19, 3309–3324.

Vaupel, J.W., Manton K.G. and Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16, 439–454.

Xue, X., 1998. Multivariate survival data under bivariate frailty: an estimating equation approach. Biometrics 54, 1631–1637.

Xue X. and Brookmeyer, R., 1996. Bivariate frailty model for the analysis of multivariate survival time. Lifetime Data Analysis 2, 277–289.

Xue X. and Ding, Y., 1999. Assessing heterogeneity and correlation of paired failure times with the bivariate frailty model. Statistics in Medicine 18, 907–918.